DOCUMENT RESUME

ED 035 057 CG 004 888

AUTHOR	Weiss, David J.
TITLE	Individualized Assessment of Differential Abilities.
INSTITUTION	American Psychological Association, Washington, D.C.; Minnesota Univ., Minneapolis.
SPONS AGENCY	Social and Rehabilitation Service (DHEW), Washington, D.C.
REPORT NO	RR-29
PUB DATE	Sep 69
NOTE	15p.; Paper presented at American Psychological Association Convention, Washington, D. C., August 31-September 4, 1969
EDRS PRICE	EDRS Price MF-$0.25 HC-$0.85
DESCRIPTORS	*Computer Oriented Programs, *Individual Tests, Measurement Instruments, Motivation, *Psychological Testing, Reliability, *Standardized Tests, Testing, *Testing Problems, Test Reliability, Tests

ABSTRACT

Today's psychological measurement depends almost exclusively on the "standardized test." A certain amount of non-standardization, however, exists in the administration of any standardized test, with the amount unknown for any given test score. Time limits on tests pose a bigger problem since another variable is introduced, pressure. Test taking motivation must also be considered. The test could be too easy or too difficult, thus boring or frustrating the individual. Reliability is also a difficulty, since there is no true reliability computed for an individual. Proper application of computer technology permits a solution to many of the problems raised by standardized tests. The tests would be individualized, with items of known difficulty grouped or stratified by level of difficulty. The testing situation could be tailored to fit an individual's preferences and/or abilities and disabilities. Administrative fluctuations and test taking motivation could be eliminated. Individualized item sequence would tailor the test to the individual, as far as difficulty is concerned. Through the item sequence, reliability would become more accurate, as the computer could more exactly pinpoint levels of difficulty. (KJ)

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ED035057

Individualized Assessment of Differential Abilities

David J. Weiss

Research Report No. 29
Work Adjustment Project
University of Minnesota
September 1969

Paper presented at symposium
"Computer-based assessment--implications
for psychometric theory and practice"
77th Annual Convention of the
American Psychological Association
Division 5

CG004888

Individualized Assessment of Differential Abilities

David J. Weiss

University of Minnesota

While the field of psychometrics has made steady progress in many technical aspects during the last thirty years or so, until now there has been no major technological innovation which has promise of freeing psychometrics from many current problems. The recent development and expansion of computer technology promises us a method which could have profound effects on both psychometric theory and practice.

The "standardized test"

Today's psychological measurement depends almost exclusively on the "standardized test." Standardization is the process of developing a common item sequence, a standard mode of administration and in many cases, a "fair" set of time limits. The objective is to present each individual with the same set of stimulus variables, and to measure the underlying ability by comparing his responses to this structured set with the responses of other individuals. The relative ability levels of two individuals are then determined by the differences in the number of items answered correctly, as compared to some norm group.

Non-standard administration. While the idea of standardization was a major achievement toward solving some of the earlier problems in psychological measurement, it, in turn, raised many other problems which have rarely been confronted. First, while the physical stimulus complex represented by the test booklet was standardized, many of the other variables surrounding test administration were not. Thus,

---

This paper was supported, in part, by Research Grant RD-1613-G from the Social and Rehabilitation Service, U.S. Department of Health, Education and Welfare.

differences in test administrator, in terms of administrator's sex, race, or committment to the task could not be standardized. Research on these variables frequently shows differences in group means when groups are tested by different administrators. It can frequently be assumed, therefore, that a certain amount of non-standardization exists in the administration of any standardized test, with the amount unknown for any given test score. Care must then be exercised in the interpretation and use of scores when conditions of administration are not clearly specified.

Time limits. A more serious problem in the use of standardized tests is in the imposition of time limits. Tests are usually timed for convenience of administration. Most psychometricians would probably agree that a "power" test is more relevant to measuring most kinds of abilities, and that the majority of criteria that we are trying to predict from ability tests are not heavily speeded. Yet most ability tests are built with time limits that force individuals to pace themselves unnecessarily. Time limits may penalize the slower, but more accurate and capable individual, while benefitting the faster individual who may have less of the ability being measured than the slower, more methodical, person. We have tried to correct for this situation in at least two ways: first, we develop time limits that permit 95% of the examinees to finish the test. This procedure is a relatively good solution, but it still may penalize the 5% of individuals who do not finish. In addition, it brings into the measurement situation extraneous personality variables which are unrelated to the ability being measured. Such variables include the ability to work under time pressure and the propensity to react to time pressure

stimulation. While perhaps there are few important differences in these two personality characteristics within the white middle class population, application of timed tests to other groups, such as the American Indian, the Black American, or the Oriental American many yield results that are invalid because of cultural differences in the propensity to react to time pressures as a source of motivation.

We have also tried to correct for the differential influence of time pressures by adopting a "correction for guessing." In this way, the individual who reacts quickly to test items and guesses freely is penalized by adjusting his score downward on the basis of the number of items he answered correctly or incorrectly. Even though the different adjustment formulas yield different results, the procedure has some merit for the fast individual who tends to guess. But we still have not found a way to perform an upward correction on the scores of the slower individual who does not guess. Granted, the differences between the fast and slow individual are reduced by the correction for guessing, but we still have no way of knowing how capable the slow individual really is. Our only solution is a completely non-speeded test, but in many cases such a measurement procedure is virtually impossible. This is especially true in the vocational assessment procedure where we wish to measure an individual's capacities on as many as ten or more independent abilities. The use of pure power tests would probably require several days of testing time for each person, with the consequent reduction in test-taking motivation that is likely to result from intensive testing.

Test-taking motivation. This leads us to the third major problem raised by the use of standardized tests. Because the

standardized test uses a common item sequence for every individual, it is wasteful of an individual's time. But, more important, it can have unmeasurable effects on test scores by affecting the individual's test-taking motivation. For some individuals, the standardized test is too easy. As a result, these individuals may get bored with answering a series of items which are not challenging enough to motivate them to high performance. The test may then be considered a "stupid waste of time" with consequent effects on the person's test scores. These invalid scores may then be used in institutional decisions relative to that individual with consequent detrimental effects on his freedom of choice.

For other individuals, a standardized test may be too difficult. In this case the individual is presented with a series of items that he may find utterly frustrating. While parts of the test may include easier items which he could validly answer, the individual may develop a negative attitude toward the test and cease answering out of frustration. While there are obviously individual differences in the reactions to such a situation, at the present time we have no way to separate those individuals who obtain a low score due to frustration from those whose low score is the result simply of lack of ability on the dimension measured by the test. While various test design formats have been developed which have implications for this problem, such as the "spiral omnibus" form of item arrangement, presentation of a set of items in a given order on a printed page is no guarantee that an individual will answer them in that order, nor can we gauge his reactions to the total stimulus complex, in order to estimate its effects on his test scores.

Reliability. A related problem that arises in the use of standardized tests is the interpretation of the reliability of the test score, or the accuracy or precision that can be attached to a given measurement. Since the standardized test is frequently built for maximum discrimination in the middle ranges of the variable being measured, measurements at the extreme are usually of differing reliability than those in the middle range of ability. This results partially from the fact that there are more items at the middle ranges; hence, the accuracy of measurement is greatest as a result of the larger number of items. At the high and low extremes there are frequently fewer items, thus reducing accuracy. The lowered accuracy at the extremes based simply on the number of items is, of course, further complicated by some of the other variables previously mentioned.

While the nature of the standardized test thus affects the accuracy or precision of measurement, we also have problems in the estimation of reliability. The most relevant estimate of reliability of an individual's test score is an error band around the score showing the degree of confidence that can be placed in interpretation and use of the obtained measurement. Reliability theory tells us to compute the standard error of measurement and use some function of it as our "error band." But the standard error of measurement is a group statistic. As such its value varies as a function of the reliability coefficient for a group of individuals, and the standard deviation of the test scores for that group. The obtained standard error of measurement figure is therefore applicable to the average member of a given group. We know that both reliability coefficients and standard deviations of test scores vary from group to group. We constantly tell our students that "there is no one reliability for a test", that

"reliability is for a given measurement." But, if one day an individual is a member of one group and another day a member of another group, it is perfectly possible that two identical measurements obtained on the same test for that individual will have different accuracy, as measured by the standard error of measurement. This may result simply because of different reliabilities and/or variabilities for the two groups. Yet there is no way to know which "error band" is the true one for that individual. Thus, the standardized test has given us a method for estimating reliability which does not permit an accurate reliability estimate of a single measurement taken on one individual; rather the accuracy of a given measurement varies artificially a result of the group an individual happens to be tested with.

The computer-based assessment system

Proper application of computer technology permits a solution to many of these problems raised by the use of standardized tests. By appropriate redesign of our testing instruments, computers can be programmed to administer psychological tests in an "individualized" or "tailored" (Lord, 1968, 1969) fashion[1]. Rather than requiring the individual to adapt himself to a standardized test, the computer can be programmed to adapt the tests to the characteristics of the individual, or to "individualize" the testing procedure.

The basic system. Individualized assessment assumes a large item pool for each ability to be measured. The optimal system would require items to be of known difficulty level, with items grouped or stratified according to difficulty level. At each difficulty level there would

---

[1]Computer-based assessment procedures have also been referred to as "branching" tests (Bayroff and Seeley, 1967), programmed tests (Cleary, Linn and Rock, 1968; Linn, Rock and Cleary, 1969) and"sequential" tests (Cronbach and Gleser, 1965).

be as many as 25 or 50 or more items in storage, each item known to be measuring the same variable at the same level. These items are then stored in the computer and identified by both dimension and difficulty level. The items are accessed by a program which controls the testing procedure.

The individual to be tested may appear for testing at any time the computer is free. He sits down at a control panel which can include a variety of input-output devices. This variety of devices for presenting items and recording responses is the first major advantage of this type of measurement.

Individualized input. The typical individualized measurement system will likely have items presented on a cathode ray tube or CRT. In many cases the individual will respond by touching a light-pen to the correct answer. But for some people, such as the physically disabled with motor problems or eye-hand coordination problems, responses may be recorded on a typewriter unit, a series of foot pedals, large push-buttons, or, within the next few years, by a computer-driven voice-writer. For those people with visual problems, the items can be presented double-sized on the cathode ray screen, or projected via computer-driven slide projectors onto movie screens in two foot letters. For the totally blind, items could be presented aurally, via computer-driven random access tape recorders. A variety of other input-output devices could be developed to tailor the testing situation to the individual's preferences and/or abilities or disabilities. Input-output devices can also be varied to maintain an interest in the testing procedure; such flexibility might be particularly effective with children.

The choice of communication devices as well as subsequent choice of items can be under computer control. Given the computer's capacity to store relevant information on an individual, the individual would simply be required to input his name or identification number to begin the testing process. All subsequent decisions based on this piece of information could be under program control. For example, our system will not require an individual to complete any one test at one sitting. The subject will be permitted to leave the testing room at any time and return at any time. Input of his name at subsequent sessions allows the computer to immediately "recall" all previous responses and to continue exactly at the point at which he left off, whether the interval be three minutes or three months. In this way, we will not force an individual to take a test under non-optimal conditions of health and/or motivation. In this way we can eliminate administrator effects and possibly reduce fluctations in scores resulting from some aspects of test-taking motivation as well as other factors usually affecting measurement accuracy.

Individualized item sequences. The individualized assessment system can be designed to eliminate or minimize many of the other problems resulting from the use of standardized tests. The test administration program can be designed to start every individual at an item of middle range difficulty or at some estimated ability level, based on other information available prior to testing. If the individual gets the item correct, the next item to be presented will be one of higher difficulty. Each correct response leads to a more difficult item, until a wrong response occurs. At that point, the program then chooses an item lower in difficulty than the lowest

item answered correctly. The effect of this procedure is to keep the test at a relevant level for the individual. If this item is answered correctly, the computer can proceed to items of more difficulty as before. Or, it can be programmed to alternate between difficult and easy items in some systematic fashion. There are an endless variety of procedures to follow at this point. The objective, of course, is to maintain the individual's interest in the task and motivation to continue, by presenting items which are not too easy for him nor too difficult.

The objective would be similar to that developed by Binet 60 years ago. We use the computer to find the level of difficulty at which the individual gets all the items correct, and the level of difficulty at which he gets all the items wrong. Having found this "lower shelf" and "upper shelf" we then, in some systematic fashion fill in the spaces between in an attempt to pinpoint the individual's capacity in terms of highest difficulty possible for him. We differ from the original Binet procedure in that we are measuring on a unidimensional variable, rather than on undefined global "intelligence", and that the items are presented by computer rather than a human psychometrist, with all the attendant interpersonal contamination factors.

Individualized precision of measurement. Given the fact that we have in storage a large number of items at each difficulty level, following identification of the individual's upper and lower shelves, we can then concentrate all further measurement within the area that is relevant for that individual. At this point we have an opportunity to clarify reliability of measurement for one individual. While the upper and lower shelves give us a gross estimate of the maximum level of difficulty of which the individual is capable, the computer can then present items at the levels in

between to obtain a revised, more accurate, maximum level within certain limits of accuracy. The reliability or precision of the obtained measurements can be controlled by the investigator and varied according to the purpose for which the measurement is to be obtained.

Under this system of measurement, a person's score on the test is the maximum difficulty level reached within a certain probability of error. If there were fifty items at a given difficulty level, with five choices each, we could assume that ten of these would be answered correctly purely by chance. Given the fact that an individual answers 25 of these 50 correctly at difficulty level X, and that he answers 8 of 50 correctly at difficulty level X+1, and further that he answers 42 of 50 correctly at difficulty level X-1, we can be fairly confident that his "true" ability level lies at difficulty level X. We can be even more confident in the accuracy of that score by presenting larger numbers of items at each of the relevant levels. This procedure permits us to narrow our lower and upper shelves to converge on the individual's ability level within the required degree of accuracy. This same process can be repeated on other ability dimensions for the same individual, with the degree of accuracy on that dimension relevant to the decision to be made on that piece of information.

The individualized assessment system would also permit us to develop and use tests of varying levels of accuracy, in terms of how finely separated the component difficulty levels are. We could develop some gross screening-type instruments for measuring second- or third-order abilities, then proceed to the finer measuring

instruments within only those gross levels that are indentified as high or low for each individual. This would permit us to measure a wide variety of abilities on each individual in a minimum amount of time to pre-determined levels of accuracy.

Other advantages of individualized meaurement. There are other aspects of an individualized system which hold promise for applications of psychometrics. Primary among these is the ability of the system to increase "motivation" by appropriate methods of feedback. The computer can inform an individual, in a variety of ways, whether his answer was right or wrong. It can use flashing lights, printed words, verbal reinforcement or food pellets dropped down a slide. We can increase motivation by tailoring the reinforcement to the individual; not all individuals are reinforced by knowledge of results. Individuals from different sub-cultures may require different kinds of reinforcing stimuli. For some, such as children, food or food-chip equivalents might be relevant. For some individuals a form of reinforcement may occur by varying systematically the inter-mix of items from various ability domains. Other ways of motivating individuals will undoubtedly by developed as individualized measurement systems become operational.

Computer-based assessment systems are obviously not bound to time limits. Rather, the computer can record the amount of time it takes for an individual to answer a given item. This information, in conjunction with information on whether each response was correct can be combined into indices of reliability. In addition, it would seem possible that judicious use of time latency measures could assist in separating out responses that are "guessed" vs. those that are obviously known immediately by the individual. Such

information can assist in helping to norm items, as well as in interpreting the results of computer-based assessment.

An additional possibility is the use of the computer to, finally, measure the ability to learn, or what we have called "aptitude" all these years. We can do this by measuring an individual's status on a given ability, then present a learning situation relevent to that ability (including knowledge of results), then finally a post-test. The difference between pre-test and post-test, with interim learning held constant, could be a measure of "ability to learn" on a given dimension. Such a procedure would seem to permit us to separate ability (measured status at one point in time) from aptitude (capacity to profit from learning). This approach would seem to have primary relevance to measurement with the "disadvantaged".

While computer-based assessment may seem relatively expensive, the cost of computer hardware is continually decreasing. Within ten years computers are likely to be as readily available as calculating machines. Most high schools, colleges, counseling agencies, employment agencies, clinics and personnel departments will have computers available to them. The measurement systems can be designed to operate on virtually any computer and on many, can operate simultaneously with scientific and business data processing. Computer-based assessment has promise for helping use solve some of the importnat problems in psychological measurement. A good system of psychometrics can assist in the solution of many of the problems of today's society, particularily in the identification of new sources of talent in untapped segments of our society.

Such a system is now under development at the Work Adjustment Project of the University of Minnesota. We will continue to explore

its implications for both the theory and practice of measurement as well as its implications for some of the vocationally-relevant problems of society.

References

Bayroff, A.G. and Seeley, L.C. An exploratory study of branching tests. Technical Report Number 188, June 1967, U.S. Army Behavioral Sciences Research Laboratory, Washington, D.C.

Cleary, T.A., Linn, R.L. and Rock, D.A. An exploratory study of programmed tests. Educational and Psychological Measurement, 1968, 28, 345-360.

Cronbach, L.L. and Gleser, G.C. Psychological tests and personnel decisions. (Second edition). Urbana, University of Illinois Press, 1965.

Linn, R.L., Rock, D.A., and Cleary, T.A. The development and evaluation of several programmed testing methods. Educational and Psychological Measurement, 1969, 29, 129-146.

Lord, F.M. Same test theory for tailored testing. Research Bulletin 68-38. Princeton, New Jersey: Educational Testing Service, 1968.

Lord, F.M. Robbins-Munro procedures for tailored testing. Research Bulletin 69-18, Princeton, New Jersey: Educational Testing Service, 1969.